**1. Introduction**

**General overview**

**Research question of interest and rational for**

When people severed in army force, they may experience excessive physical exercise. We want to see if there is a difference between citizens and veterans in the incidence of Heart Attack across all other influencing factors, and how different they are.

**Review on existing research**

**2. Data**

**Data set, source, general information & context**

The data comes from Behavioral Risk Factor Surveillance System (BRFSS) 2016 Data set. BRFSS is the largest continuously conducted health survey system in the word since it completes more than 400,000 adult through telephone interview every year. BRFSS collects health-related data from residents across U.S., including their health-related risk behaviors, chronic health conditions, and use of preventive services.

**Sample size, variables, results on first descriptive and numeric analysis**

In the BRFSS 2016 Data Set, there are 63,919 veterans and 421,398 citizens. We pick up Heart Attack as a response variable and 6 variables as people’s background information, including Education, Age, Income, Sex, Marital Status, and Employment Status. After the missing values are moved out of the data sample, there are 49,222 veterans and 316,093 citizens.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| level | Sex | Veteran (1) | | Citizen (2) | |
| 1 | Male | 44794 | 91.00% | 115666 | 36.59% |
| 2 | Female | 4428 | 9.00% | 200427 | 63.41% |

Table 1：The distribution of Gender for Veteran and Citizen.

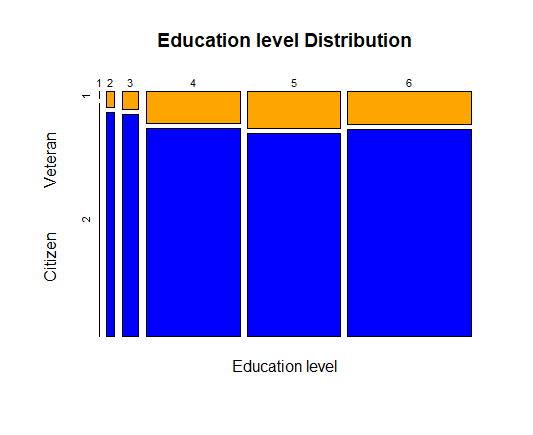
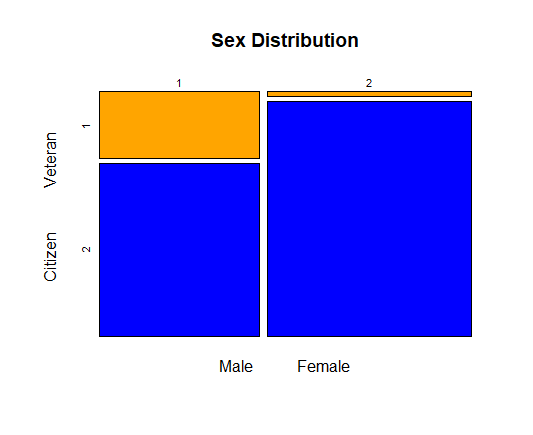


Figure 1: The distribution of Gender Figure 2: The distribution of Education

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| level | Education | Veteran | | Citizen | |
| 1 | Never attend school | 12 | 0.02% | 427 | 0.14% |
| 2 | Elementary school | 554 | 1.13% | 7909 | 2.50% |
| 3 | Some high school | 1330 | 2.70% | 16844 | 5.33% |
| 4 | High school graduate | 13562 | 27.55% | 88356 | 27.95% |
| 5 | Some college | 15366 | 31.22% | 85635 | 27.09% |
| 6 | College graduate | 18398 | 37.38% | 116922 | 36.99% |

Table 2：The distribution of Education for Veteran and Citizen.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| level | Age | Veteran | | Citizen | |
| 1 | 18<=Age<=24 | 772 | 1.57% | 18779 | 5.94% |
| 2 | 25<=Age<=34 | 2375 | 4.83% | 33942 | 10.74% |
| 3 | 35<=Age<=44 | 2895 | 5.88% | 38720 | 12.25% |
| 4 | 45<=Age<=54 | 5595 | 11.37% | 52630 | 16.65% |
| 5 | 55<=Age<=64 | 8584 | 17.44% | 73079 | 23.12% |
| 6 | 65<=Age | 29001 | 58.92% | 98943 | 31.30% |

Table 3：The distribution of Age for Veteran and Citizen.

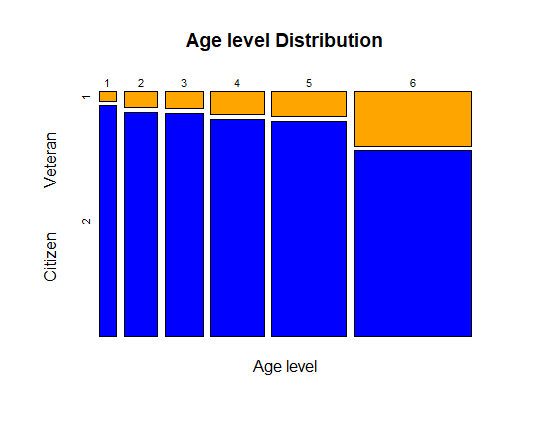
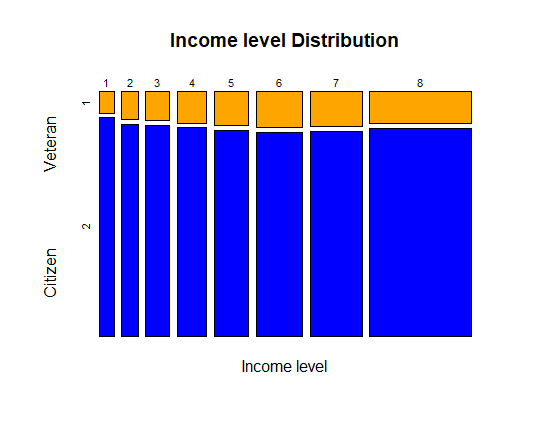
 

Figure 3: The distribution of Age Figure 4: The distribution of Income

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| level | Income/year | Veteran | | Citizen | |
| 1 | Income<10k | 1514 | 3.08% | 15472 | 4.89% |
| 2 | 10k<=income<15k | 2224 | 4.52% | 17066 | 5.40% |
| 3 | 15k<=income<20k | 3302 | 6.71% | 23746 | 7.51% |
| 4 | 20k<=income<25k | 4390 | 8.92% | 29117 | 9.21% |
| 5 | 25k<=income<35k | 5621 | 11.42% | 33879 | 10.72% |
| 6 | 35k<=income<50k | 7929 | 16.11% | 44828 | 14.18% |
| 7 | 50k<=income<75k | 8634 | 17.54% | 50618 | 16.01% |
| 8 | 75k<=income | 15608 | 31.71% | 101367 | 32.07% |

Table 4：The distribution of Income for Veteran and Citizen.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| level | Marital Status | Veteran | | Citizen | |
| 1 | Married | 28231 | 57.35% | 167568 | 53.01% |
| 2 | Divorced | 7069 | 14.36% | 45184 | 14.29% |
| 3 | Widowed | 5541 | 11.26% | 37591 | 11.89% |
| 4 | Separated | 894 | 1.82% | 6769 | 2.14% |
| 5 | Never married | 6175 | 12.55% | 48759 | 15.43% |
| 6 | Has unmarried life partner | 1312 | 2.67% | 10222 | 3.23% |

Table 5：The distribution of Marital Status for Veteran and Citizen.

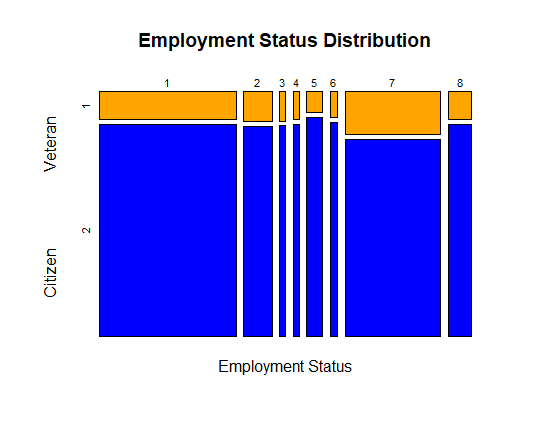
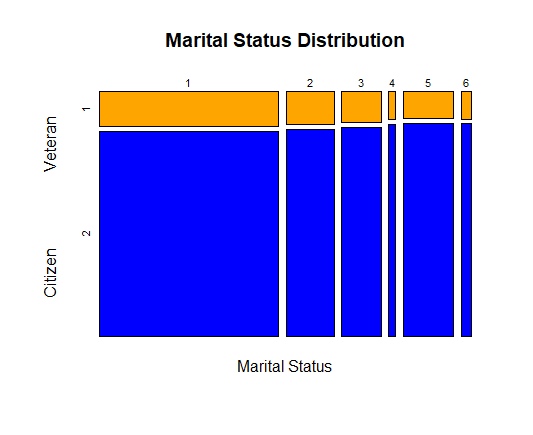


Figure 5: The distribution of Marital Status Figure 6: The distribution of Employment

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| level | Employment Status | Veteran | | Citizen | |
| 1 | Employed | 18344 | 37.27% | 138044 | 43.67% |
| 2 | Self-employed | 4137 | 8.40% | 28788 | 9.11% |
| 3 | Out of work for 1 year or more | 847 | 1.72% | 6070 | 1.92% |
| 4 | Out of work for less than 1 year | 807 | 1.64% | 6045 | 1.91% |
| 5 | Homemaker | 1604 | 3.26% | 16509 | 5.22% |
| 6 | Student | 911 | 1.85% | 7450 | 2.36% |
| 7 | Retired | 19471 | 39.56% | 89442 | 28.30% |
| 8 | Unable to work | 3101 | 6.30% | 23745 | 7.51% |

Table 6：The distribution of Employment Status for Veteran and Citizen.

There are some difference in the distributions of veterans and citizens in some variables. Citizens have 63.41% female, while veterans only have 9%; 58.92% of veterans are 55 years old or older, while only 31.30% citizens are in this age level. Due to the age distribution, 39.56% of veterans are retired, which is 11.26% higher than citizens.

In the data sample, there are 6468 veterans (13.14%) and 16035 citizens (5.07%) who have had Heart Attack. The incidence rate of veterans is significantly higher than it of citizens (p-value<2.2e-16). When we look into each single variables, the incidence rate of veterans is higher for every level.

|  |  |  |  |
| --- | --- | --- | --- |
| level | Heart Attack | Veteran | Citizen |
| 1 | Have had before | 6468 | 16035 |
| 2 | Never have | 42754 | 300058 |

Table 7：The distribution of Heart Attack for Veteran and Citizen.

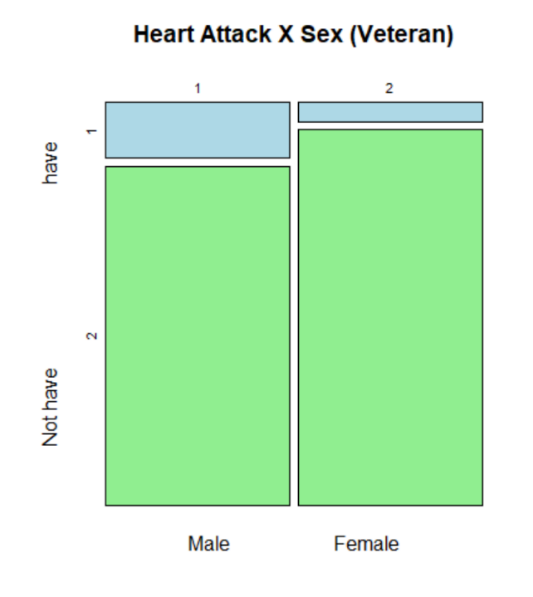


Figure 7: The incidence rate in different Gender levels.

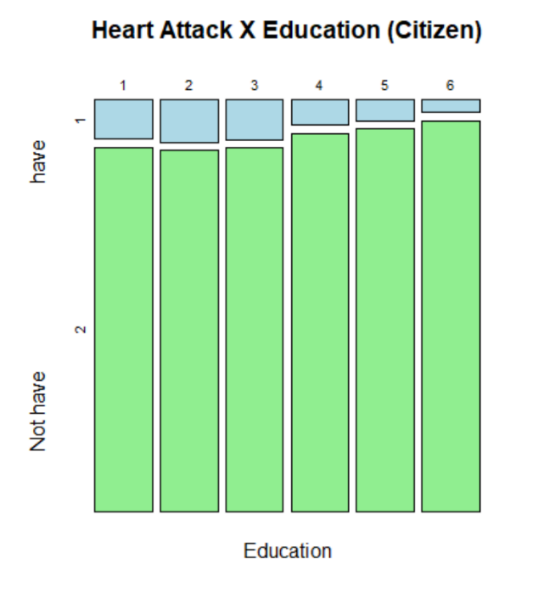
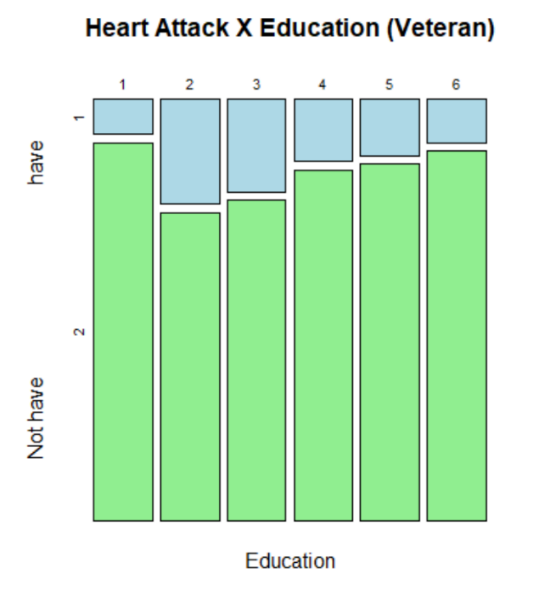


Figure 8: The incidence rate in different Education levels.

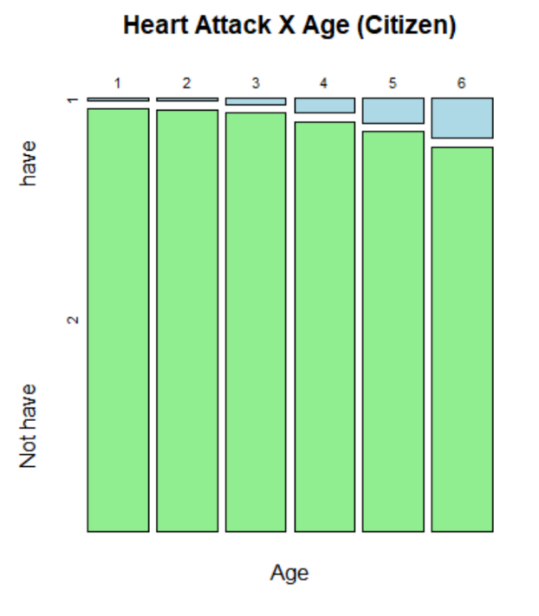
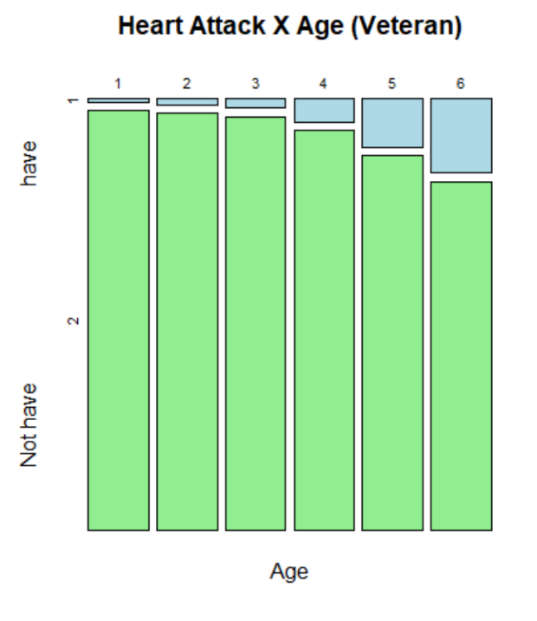


Figure 8: The incidence rate in different Age levels.

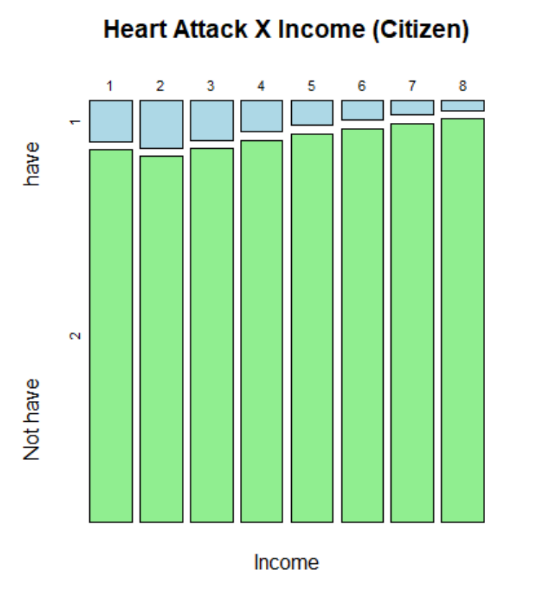
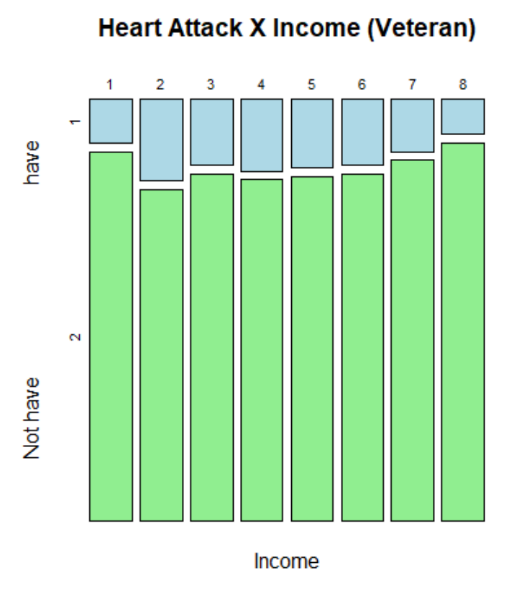


Figure 9: The incidence rate in different Income levels.

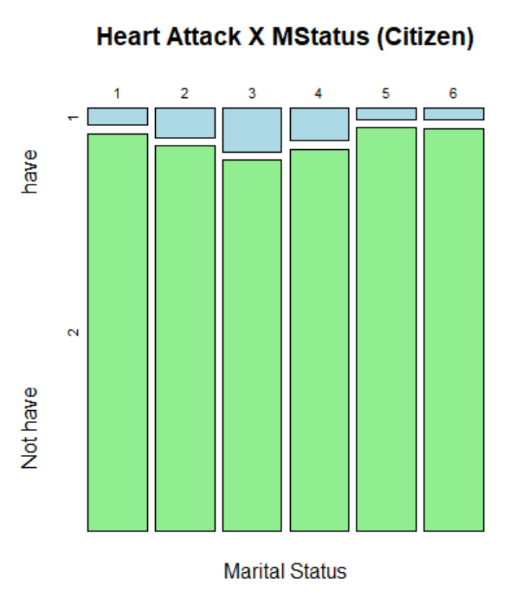
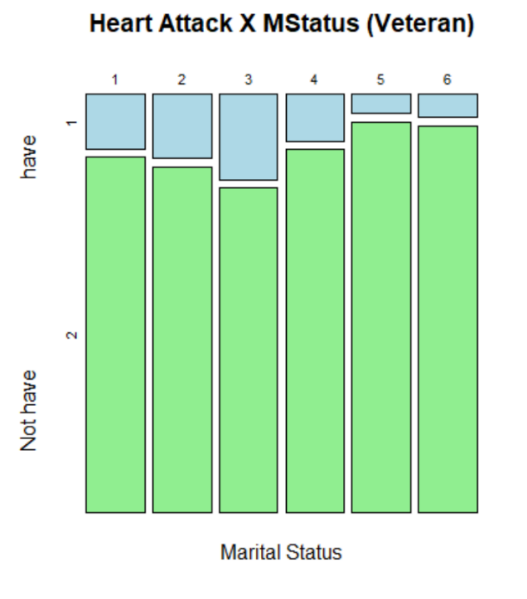


Figure 10: The incidence rate in different Marital status.

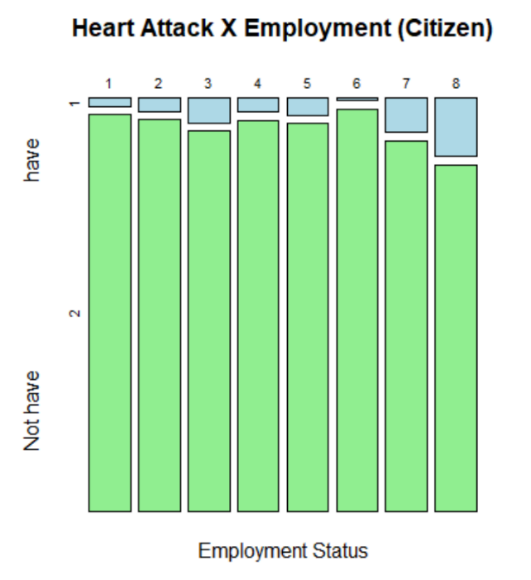
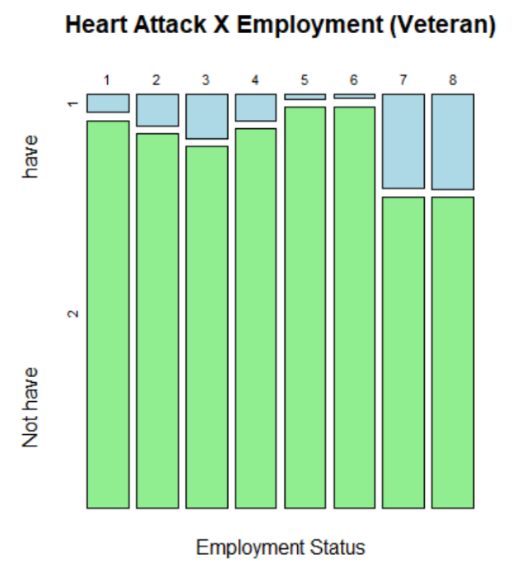


Figure 11: The incidence rate in different Employment Status.

Because of the extreme differences between veterans and citizens on the distribution of Age, Sex, and Employment status, the illness incidence rates are weighted by the population weights.

|  |  |  |  |
| --- | --- | --- | --- |
| Age level | veteran.inc | citizen.inc | weight |
| 1 | 0.0065 | 0.0034 | 0.1266 |
| 2 | 0.0135 | 0.0064 | 0.1738 |
| 3 | 0.0214 | 0.0138 | 0.1634 |
| 4 | 0.0545 | 0.0341 | 0.1694 |
| 5 | 0.1134 | 0.0574 | 0.1666 |
| 6 | 0.1755 | 0.0932 | 0.2002 |
| Adjusted incidence in total | 0.0699 | 0.0378 | ---- |

Table x: The weighted incidence rate with Age weights while all other variables combined.

|  |  |  |  |
| --- | --- | --- | --- |
| Employment level | veteran.incidence | citizen.incidence | weight |
| 1 | 0.0424 | 0.0203 | 0.4774 |
| 2 | 0.0761 | 0.0329 | 0.0874 |
| 3 | 0.1074 | 0.0600 | 0.0252 |
| 4 | 0.0632 | 0.0344 | 0.0282 |
| 5 | 0.0100 | 0.0420 | 0.0642 |
| 6 | 0.0088 | 0.0050 | 0.0574 |
| 7 | 0.2304 | 0.0846 | 0.1797 |
| 8 | 0.2332 | 0.1437 | 0.0695 |
| Adjusted incidence in total | 0.0901 | 0.0432 | ---- |

Table x: The weighted incidence rate with Employment status weights while all other variables combined.

|  |  |  |  |
| --- | --- | --- | --- |
| Sex level | veteran.incidence | citizen.incidence | weight |
| 1 | 0.1396 | 0.0639 | 0.4866 |
| 2 | 0.0483 | 0.0431 | 0.5133 |
| Adjusted incidence in total | 0.0927 | 0.0532 | ---- |

Table x: The weighted incidence rate with Gender weights while all other variables combined.

However, we could not conclude that veterans have a higher incidence of Heart Attack than citizens because there are much more elderly people in veterans. The higher incidence may be caused by too many elderly people, or the incidence in veterans is higher than in citizens only in a specific group.

**3. Methodology**

**3-1 the Detection of the difference**

Apparently, the Heart Attack incidence rates of Veterans and Citizens are different based on the analysis of the raw data. However, the population structures of the two groups are totally different. Rubin-Casual Model is used to detect if there is a difference between two groups on the incidence rates.

The matching method is perfect matching instead of score matching because of the abundant samples. Within each block (with exactly same factor levels), the same numbers of citizens and veterans were randomly selected. The selecting amount in each block was set to be as large as possible. In total, 42,161 citizens and 42,161 veterans are selected over all blocks. The propensity scores of the two groups are the same due to the matching is the perfect matching (same joint distributions of all six factors in each block). After repeating the random selection with random seeds of 1 to 75, the Heart Attack incidence rates of the two groups were obtained.

For obtaining a more generalized result, the American population weights (provided by the raw data) are used to deter the generalized different on incidence rate. The incidence rates of veterans and citizens in each block are calculated first, and the block is ignored if the incidence rate is 0 or 1 (which means there is only one heart-attack type in this block). Only 1018 blocks left for Veterans, and 3337 blocks left for Citizens. Then, the average incidence rates are calculated with population weights. Finally, the difference of two incidence rates is obtained and tested to see if it is significant.

**3-2 The reason of the difference**

In the data, several category variables are collinear, Education and Income, Age and Employment, for example. In this case, Principal Component Analysis helps to detect the specific factor levels which relate to high incidence rate. Also, the possible interaction effects between variables are noticed from the image analysis.

Six factors (including Sex, Education, Employment status, Income, Marital status and Age) are used in PCA (with the base package in R vision 3.4.3) after standardization. The result is shown separately for Veterans and Citizens.

After the PCA, a logistic model is built to predict the Heart Attack incidence for each unit.

**4. Results**

**4-1 Rubin-Casual Model**

Theoretically, there are 27648 blocks in the data (2 (Sex) 6 (Education) \* 6 (Age) \* 6 (MaritalStatus) \* 8 (Employment) \* 8 (Income) =27648). However, only 15873 blocks are with units inside.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **V1** | **V0** | **NV1** | **NV0** | **V\_incidence** | **NV\_incidence** | **p\_value** |
| **1** | **4974** | **37187** | **4252** | **37909** | **0.117976** | **0.100851** | **1.80E-15** |
| **2** | **5010** | **37151** | **4270** | **37891** | **0.11883** | **0.101278** | **4.23E-16** |
| **3** | **5013** | **37148** | **4252** | **37909** | **0.118901** | **0.100851** | **5.82E-17** |
| **4** | **4953** | **37208** | **4244** | **37917** | **0.117478** | **0.100662** | **5.22E-15** |
| **5** | **4953** | **37208** | **4254** | **37907** | **0.117478** | **0.100899** | **1.28E-14** |
| **6** | **5006** | **37155** | **4265** | **37896** | **0.118735** | **0.10116** | **3.75E-16** |
| **7** | **5035** | **37126** | **4209** | **37952** | **0.119423** | **0.099832** | **9.58E-20** |
| **8** | **4991** | **37170** | **4241** | **37920** | **0.11838** | **0.100591** | **1.45E-16** |
| **9** | **5009** | **37152** | **4206** | **37955** | **0.118806** | **0.09976** | **8.57E-19** |
| **10** | **5016** | **37145** | **4255** | **37906** | **0.118973** | **0.100923** | **5.94E-17** |
| **11** | **4932** | **37229** | **4238** | **37923** | **0.11698** | **0.100519** | **1.78E-14** |
| **12** | **5000** | **37161** | **4225** | **37936** | **0.118593** | **0.100211** | **1.35E-17** |
| **13** | **4976** | **37185** | **4247** | **37914** | **0.118024** | **0.100733** | **9.55E-16** |
| **14** | **5036** | **37125** | **4177** | **37984** | **0.119447** | **0.099073** | **2.76E-21** |
| **15** | **5013** | **37148** | **4241** | **37920** | **0.118901** | **0.100591** | **1.99E-17** |
| **16** | **5033** | **37128** | **4201** | **37960** | **0.119376** | **0.099642** | **4.99E-20** |
| **17** | **5026** | **37135** | **4223** | **37938** | **0.11921** | **0.100164** | **9.74E-19** |
| **18** | **5020** | **37141** | **4205** | **37956** | **0.119067** | **0.099737** | **2.70E-19** |
| **19** | **5017** | **37144** | **4241** | **37920** | **0.118996** | **0.100591** | **1.38E-17** |
| **20** | **4955** | **37206** | **4229** | **37932** | **0.117526** | **0.100306** | **1.11E-15** |
| **21** | **5013** | **37148** | **4260** | **37901** | **0.118901** | **0.101041** | **1.26E-16** |
| **22** | **4993** | **37168** | **4242** | **37919** | **0.118427** | **0.100614** | **1.33E-16** |
| **23** | **5004** | **37157** | **4208** | **37953** | **0.118688** | **0.099808** | **1.69E-18** |
| **24** | **5027** | **37134** | **4199** | **37962** | **0.119233** | **0.099594** | **7.27E-20** |
| **25** | **4970** | **37191** | **4190** | **37971** | **0.117881** | **0.099381** | **6.63E-18** |
| **26** | **5028** | **37133** | **4195** | **37966** | **0.119257** | **0.0995** | **4.31E-20** |
| **27** | **4999** | **37162** | **4222** | **37939** | **0.118569** | **0.10014** | **1.10E-17** |
| **28** | **5000** | **37161** | **4219** | **37942** | **0.118593** | **0.100069** | **7.44E-18** |
| **29** | **5035** | **37126** | **4232** | **37929** | **0.119423** | **0.100377** | **1.04E-18** |
| **30** | **5009** | **37152** | **4281** | **37880** | **0.118806** | **0.101539** | **1.29E-15** |
| **31** | **5027** | **37134** | **4227** | **37934** | **0.119233** | **0.100259** | **1.33E-18** |
| **32** | **4977** | **37184** | **4238** | **37923** | **0.118047** | **0.100519** | **3.76E-16** |
| **33** | **4977** | **37184** | **4254** | **37907** | **0.118047** | **0.100899** | **1.68E-15** |
| **34** | **4992** | **37169** | **4212** | **37949** | **0.118403** | **0.099903** | **7.77E-18** |
| **35** | **4989** | **37172** | **4208** | **37953** | **0.118332** | **0.099808** | **6.88E-18** |
| **36** | **5057** | **37104** | **4228** | **37933** | **0.119945** | **0.100282** | **8.31E-20** |
| **37** | **4999** | **37162** | **4185** | **37976** | **0.118569** | **0.099262** | **2.54E-19** |
| **38** | **5007** | **37154** | **4196** | **37965** | **0.118759** | **0.099523** | **3.69E-19** |
| **39** | **4991** | **37170** | **4232** | **37929** | **0.11838** | **0.100377** | **6.09E-17** |
| **40** | **4960** | **37201** | **4283** | **37878** | **0.117644** | **0.101587** | **9.22E-14** |
| **41** | **4977** | **37184** | **4227** | **37934** | **0.118047** | **0.100259** | **1.32E-16** |
| **42** | **5010** | **37151** | **4219** | **37942** | **0.11883** | **0.100069** | **2.93E-18** |
| **43** | **4979** | **37182** | **4220** | **37941** | **0.118095** | **0.100093** | **5.62E-17** |
| **44** | **4966** | **37195** | **4233** | **37928** | **0.117787** | **0.100401** | **6.17E-16** |
| **45** | **4953** | **37208** | **4208** | **37953** | **0.117478** | **0.099808** | **1.82E-16** |
| **46** | **4981** | **37180** | **4224** | **37937** | **0.118142** | **0.100187** | **6.91E-17** |
| **47** | **5046** | **37115** | **4270** | **37891** | **0.119684** | **0.101278** | **1.69E-17** |
| **48** | **5031** | **37130** | **4235** | **37926** | **0.119328** | **0.100448** | **2.06E-18** |
| **49** | **4987** | **37174** | **4233** | **37928** | **0.118285** | **0.100401** | **9.61E-17** |
| **50** | **4989** | **37172** | **4221** | **37940** | **0.118332** | **0.100116** | **2.50E-17** |
| **51** | **5009** | **37152** | **4221** | **37940** | **0.118806** | **0.100116** | **3.94E-18** |
| **52** | **5014** | **37147** | **4187** | **37974** | **0.118925** | **0.09931** | **7.29E-20** |
| **53** | **4981** | **37180** | **4252** | **37909** | **0.118142** | **0.100851** | **9.86E-16** |
| **54** | **4956** | **37205** | **4219** | **37942** | **0.117549** | **0.100069** | **3.97E-16** |
| **55** | **5011** | **37150** | **4239** | **37922** | **0.118854** | **0.100543** | **1.96E-17** |
| **56** | **4999** | **37162** | **4229** | **37932** | **0.118569** | **0.100306** | **2.20E-17** |
| **57** | **4961** | **37200** | **4291** | **37870** | **0.117668** | **0.101777** | **1.69E-13** |
| **58** | **5019** | **37142** | **4286** | **37875** | **0.119044** | **0.101658** | **8.60E-16** |
| **59** | **4994** | **37167** | **4192** | **37969** | **0.118451** | **0.099428** | **8.48E-19** |
| **60** | **5010** | **37151** | **4208** | **37953** | **0.11883** | **0.099808** | **9.57E-19** |
| **61** | **4996** | **37165** | **4232** | **37929** | **0.118498** | **0.100377** | **3.87E-17** |
| **62** | **4962** | **37199** | **4216** | **37945** | **0.117692** | **0.099998** | **1.76E-16** |
| **63** | **4991** | **37170** | **4265** | **37896** | **0.11838** | **0.10116** | **1.38E-15** |
| **64** | **4997** | **37164** | **4189** | **37972** | **0.118522** | **0.099357** | **4.67E-19** |
| **65** | **5017** | **37144** | **4245** | **37916** | **0.118996** | **0.100685** | **2.05E-17** |
| **66** | **5027** | **37134** | **4245** | **37916** | **0.119233** | **0.100685** | **8.17E-18** |
| **67** | **4997** | **37164** | **4238** | **37923** | **0.118522** | **0.100519** | **6.34E-17** |
| **68** | **4989** | **37172** | **4293** | **37868** | **0.118332** | **0.101824** | **2.06E-14** |
| **69** | **4969** | **37192** | **4251** | **37910** | **0.117858** | **0.100828** | **2.53E-15** |
| **70** | **4982** | **37179** | **4207** | **37954** | **0.118166** | **0.099784** | **1.19E-17** |
| **71** | **4999** | **37162** | **4242** | **37919** | **0.118569** | **0.100614** | **7.80E-17** |
| **72** | **4953** | **37208** | **4259** | **37902** | **0.117478** | **0.101018** | **2.01E-14** |
| **73** | **4970** | **37191** | **4210** | **37951** | **0.117881** | **0.099855** | **4.79E-17** |
| **74** | **4968** | **37193** | **4232** | **37929** | **0.117834** | **0.100377** | **4.72E-16** |
| **75** | **4988** | **37173** | **4255** | **37906** | **0.118308** | **0.100923** | **7.09E-16** |

**Table X: The selected result of 75 random selections. V1 is the number of veterans who have had heart attack. NV1 is the number of citizens who have had heart attack. V0 is the number of healthy veterans. NV0 is the number of healthy citizens. P-value is for the proportion test.**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Min** | **1st Qu.** | **Median** | **Mean** | **3rd Qu.** | **Max** |
| **V\_incidence** | **0.1170** | **0.1180** | **0.1185** | **0.1185** | **0.1189** | **0.1199** |
| **NV\_incidence** | **0.0991** | **0.0999** | **0.1004** | **0.1008** | **0.1008** | **0.1018** |

**Table X: The distribution of incidence rates among 75 times repeats.**

Among the 75 times repeating, veterans’ incidences were always higher than citizens’ around 0.01 (mean incidence 0.106 for citizens and 0.118 for veterans), which is significant in proportion test. In this process, the selected data structure is more like the structure of veterans, since more than 85% of veterans were selected.

Although there is a significant difference between veterans and citizens’ incidence rates in proportion test, the difference is only 1%, which is different from what raw data say.

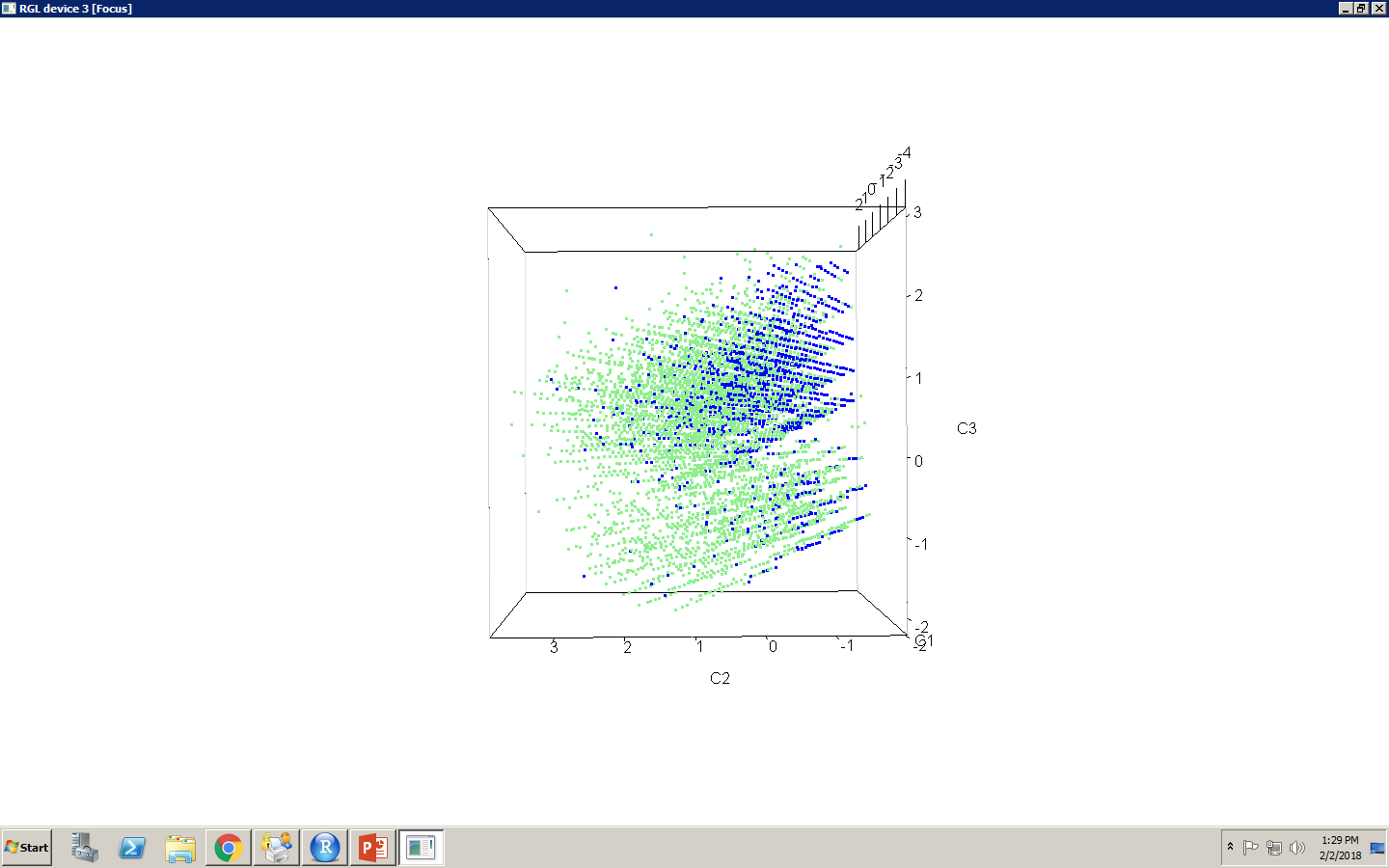
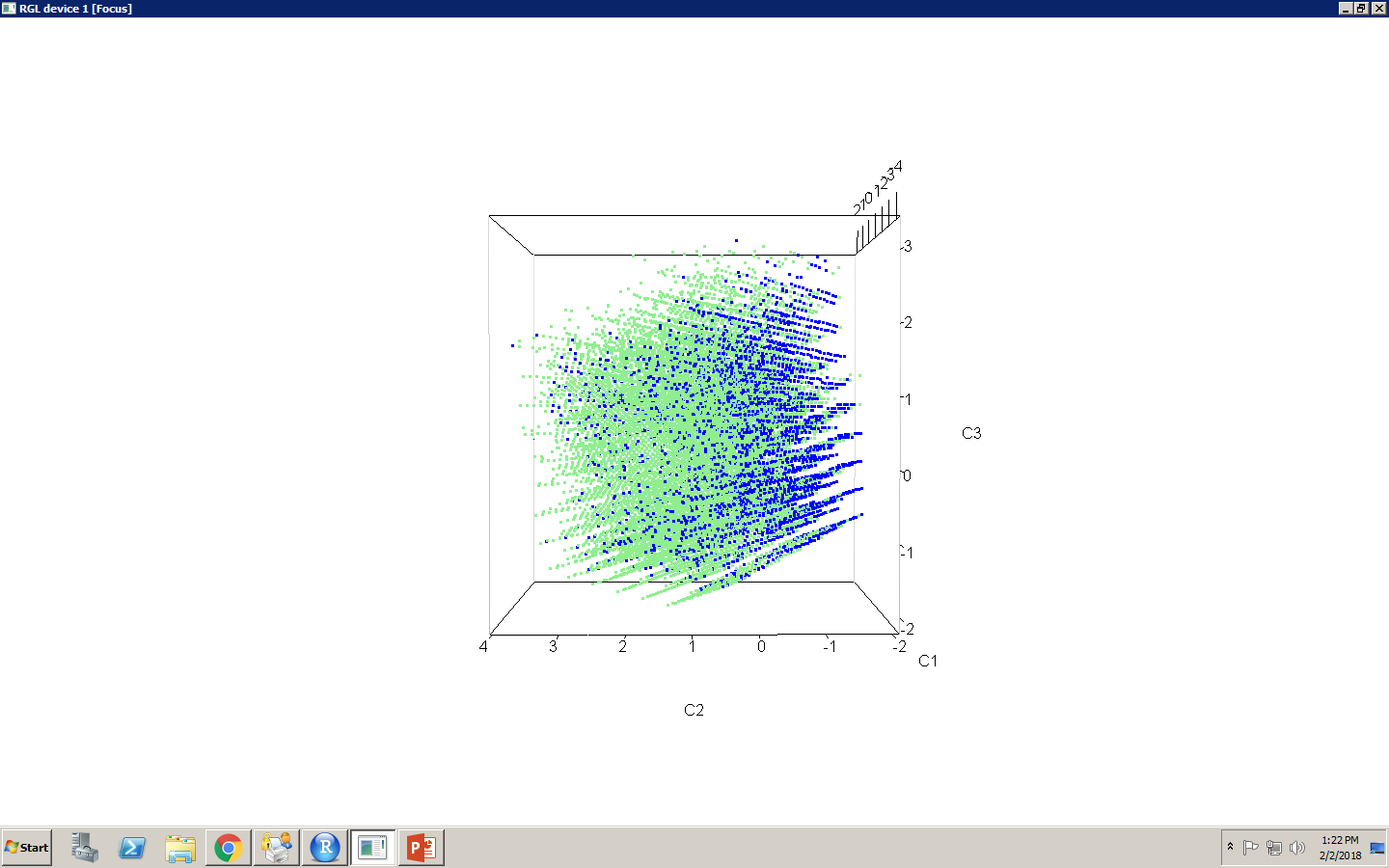
**4-2 Principal Component Analysis**

For doing PCA, the observations with any missing values were moved out from the sample, and every column was standardized. Next, columns of Education, Age, Income, Sex, Marital Status, and Employment Status were put into a PCA function in R. Then the result of PCA was shown as below.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Comp.1 | Comp.2 | Comp.3 | Comp.4 | Comp.5 | Comp.6 |
| Standard deviation | 1.2433934 | 1.0661779 | 1.0076810 | 0.9744306 | 0.9099876 | 0.72403317 |
| Proportion of Variance | 0.2576712 | 0.1894559 | 0.1692368 | 0.1582525 | 0.1380129 | 0.08737067 |
| Cumulative Proportion | 0.2576712 | 0.4471271 | 0.6163639 | 0.7746164 | 0.9126293 | 1.00000000 |

Table 8: The details of principle components.

According to the result from PCA, the first three components could explain 61.63% variation. For visual analysis, citizens and veterans were separated by their Veteran Value, and patients were colored in Blue, while others were colored in Green.



Citizen Veteran

Figure 12: The screenshot of 3D plot of units with 3 principal components. Blue represents those who have had Heart Attack.

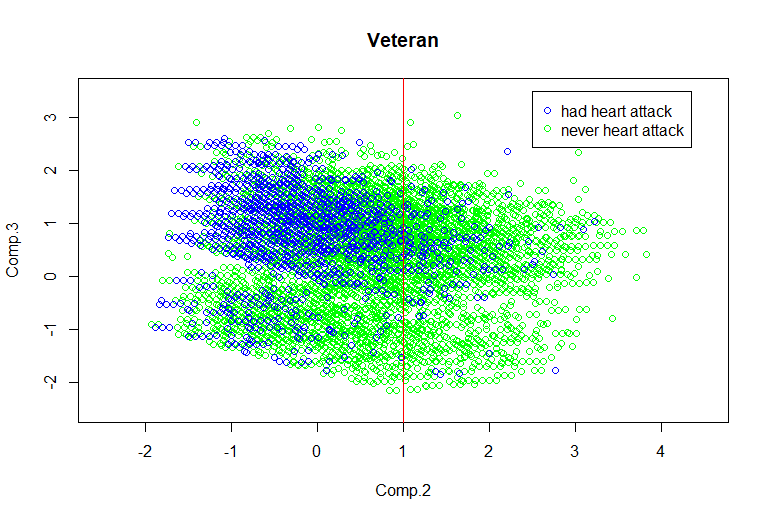
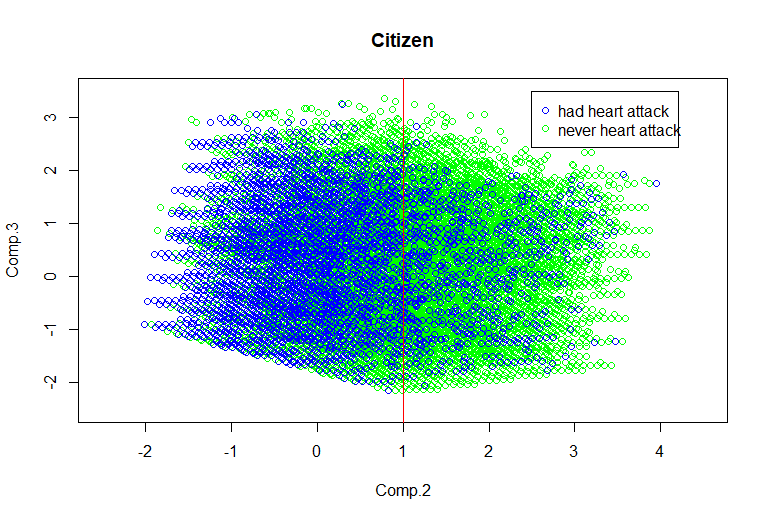


Figure 13: The 2D plot of units with 2 comp.2 and comp.3. Blue represents those who have had Heart Attack.

Apparently, the range of veteran units is a subset of the range of citizen units. It is obviously that blue units tend to have a small (or negative) component 2 value (C2) for both groups. A line of C2=1 is added on the plots due to visually analysis. In the area of C2<1, there are 88.4% of all Veterans and 81.0% of all Citizens. However, 97.7% of veteran patients and 93.81% of citizen patients are in the area of C2<1. The area of C2<1 is a higher probability area of heart attack, which is consisted with the Principal Component Logistic Regression.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Estimate | Std.Error | Z Value | Pr(>|Z|) |  |
| (Intercept) | 3.172314 | 0.024533 | 129.307 | < 2e-16 | \*\*\* |
| Veteran2 | 0.293586 | 0.022944 | 12.796 | < 2e-16 | \*\*\* |
| Comp.1 | 0.645471 | 0.007574 | 85.221 | < 2e-16 | \*\*\* |
| Comp.2 | 0.799214 | 0.011332 | 70.524 | < 2e-16 | \*\*\* |
| Comp.3 | -0.423487 | 0.009529 | -44.440 | < 2e-16 | \*\*\* |
| Comp.4 | -0.302219 | 0.009518 | -31.753 | < 2e-16 | \*\*\* |
| Comp.5 | 0.387190 | 0.012888 | 30.043 | < 2e-16 | \*\*\* |
| Comp.6 | -0.045293 | 0.012355 | -3.666 | 0.000246 | \*\*\* |

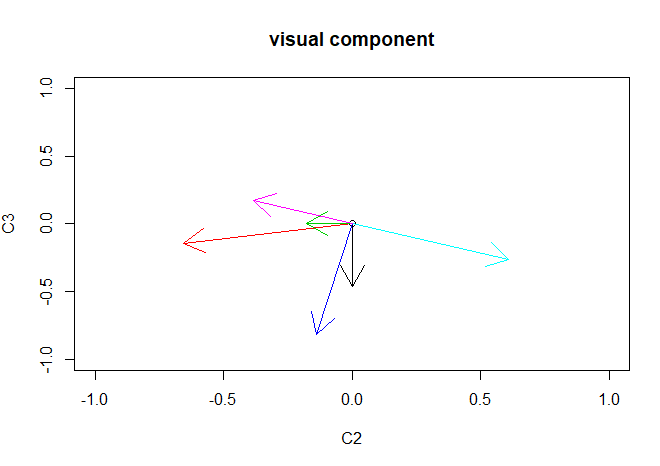
Table 9: The logistic regression results based on the factor of Veteran Experience and 6 components.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| C2 | C2<1 | | C2>1 | |
|  | Veteran | Citizen | Veteran | Citizen |
| Heart Attack=1(Yes) | 6317(14.5%) | 15042(5.9%) | 151(2.6%) | 993(1.6%) |
| Heart Attack=2(No) | 37180(85.5%) | 240851(94.1%) | 5574(97.4%) | 59207(98.4%) |

Table 10: the incidence rate for Veterans and Citizens in two areas. The Veterans group has higher rate than Citizens group in both of the areas.

|  |  |  |  |
| --- | --- | --- | --- |
| Variables | Comp.1 | Comp.2 | Comp.3 |
| Education | 0.296 |  | -0.459 |
| Age | -0.215 | -0.655 | -0.143 |
| Income | 0.662 | -0.176 |  |
| Sex | -0.124 | -0.138 | -0.819 |
| MStatus | -0.377 | 0.609 | -0.260 |
| Emloyment | -0.519 | -0.385 | 0.175 |

Table 11: the score of the first three components.



Education

Age

Income

Sex

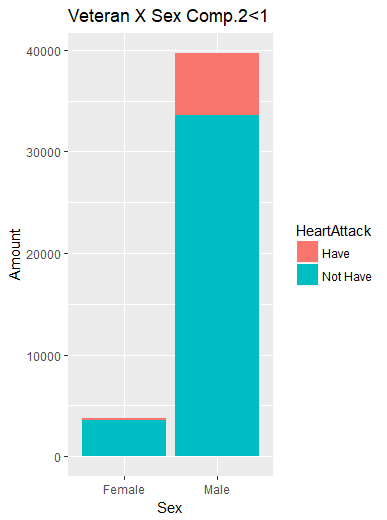
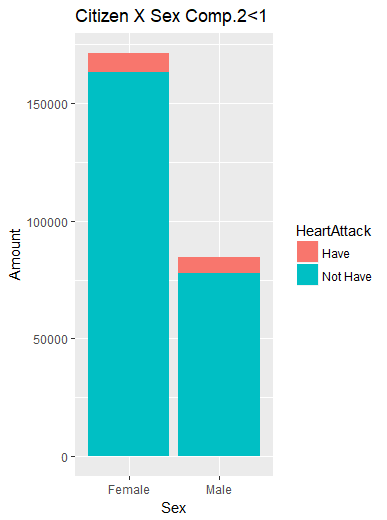
Marital Status

Employment

Figure 14: The original factors on the 2D map of comp.2 and comp.3.

According to the details of Component 2, C2<1 represents one or more characteristics among high Age, high Income, female, married, and unemployed or retired.

Then consider the interaction between factors.



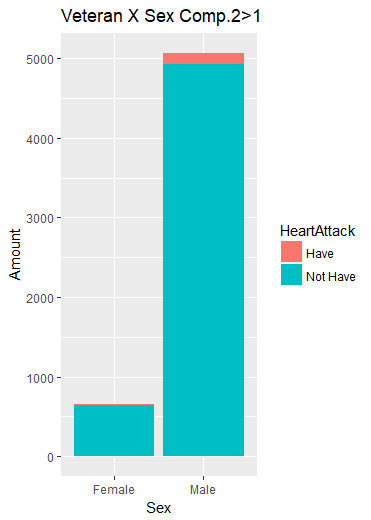
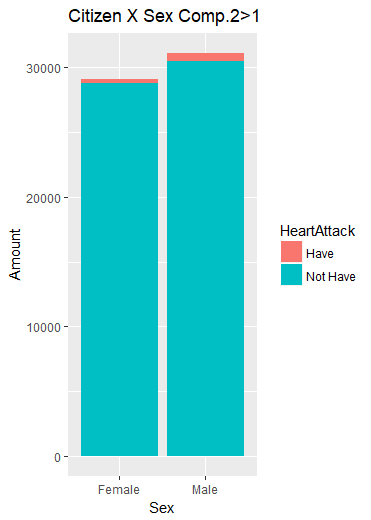


Figure 15: The distribution on Sex for the group with C2<1 (left) and for the group with C2>1(right).

Due to the Data set, the gender distribution is extremely uneven for Veterans and Citizens. Noticeable, Male Veterans have more probability to have heart attack than Female Veterans or (Male and Female) Citizens.

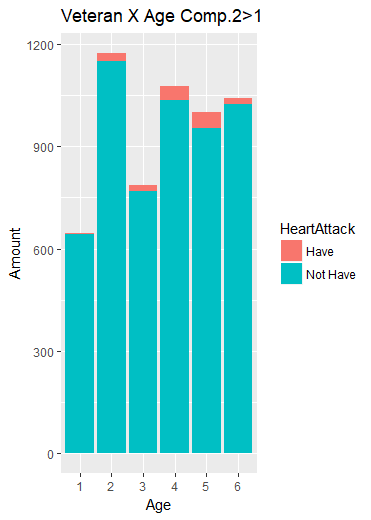
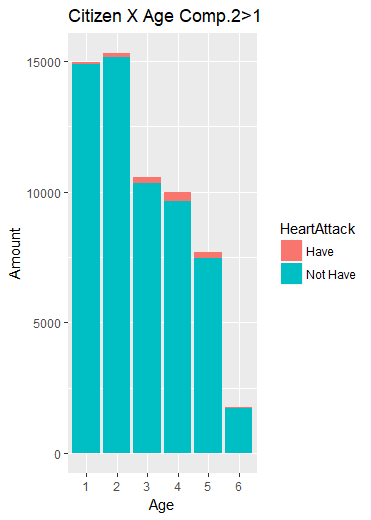
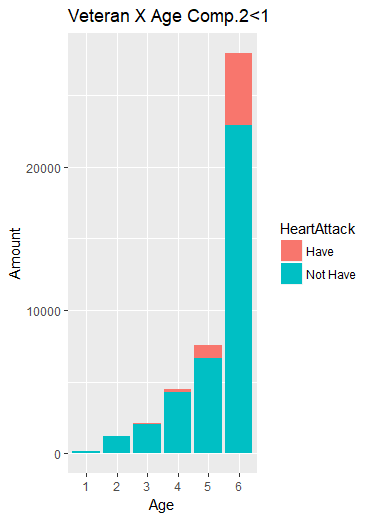
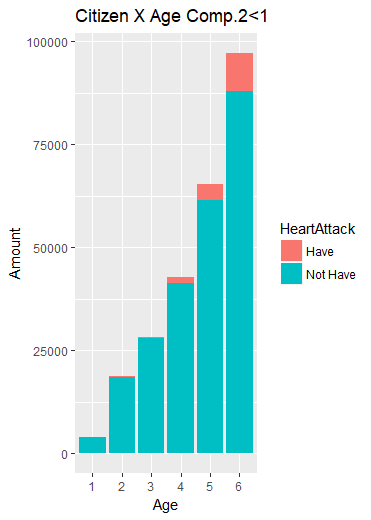


Figure 16: The distribution on Age for the group with C2<1 (left) and for the group with C2>1(right).

The distributions of the 4 plots in Figure X are totally different with each other (p-values all smaller than 2.2e-16). Overall, in the area of C2<1, Age 6 takes the largest part of Citizens and Veterans. Differently, in the area of C2>1, Citizens have less units with higher Age level, while Veterans are distributed more complicatedly. Definitely, Age is an important factor with the probability of Heart Attack, and there is an interaction effect of the factor Veteran and high Age levels.

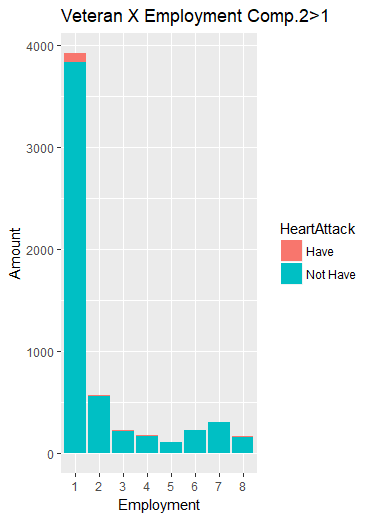
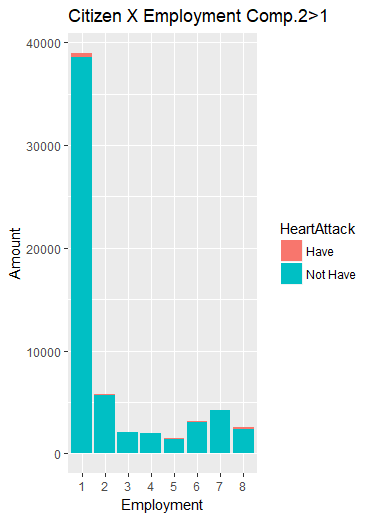
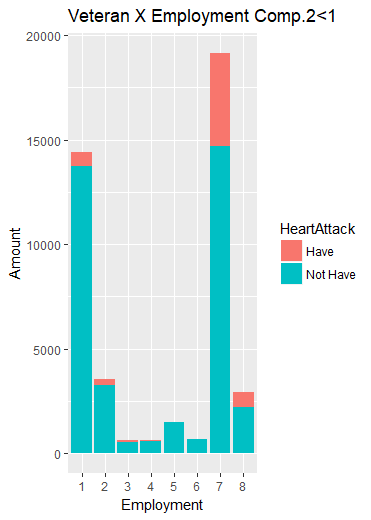
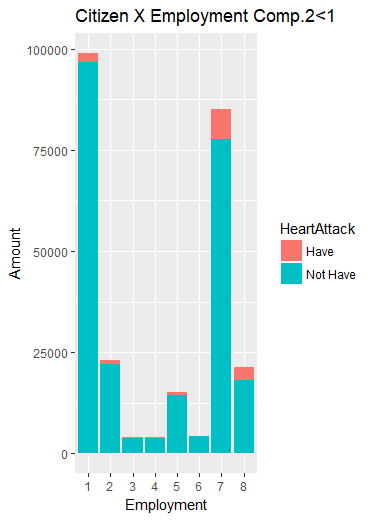


Figure 17: The distribution on Employment for the group with C2<1 (left) and for the group with C2>1(right).

Figure 17 shows the totally different distributions between the group whose Comp.2<1 and whose Comp.2>1, and the distributions of Citizens and Veterans with Comp.2>1 are similar (p-value=0.95). In the area of C2<1, Employment level 7 is the highest incidence rate level for both of Veterans (23.4%) and Citizens (8.9%).

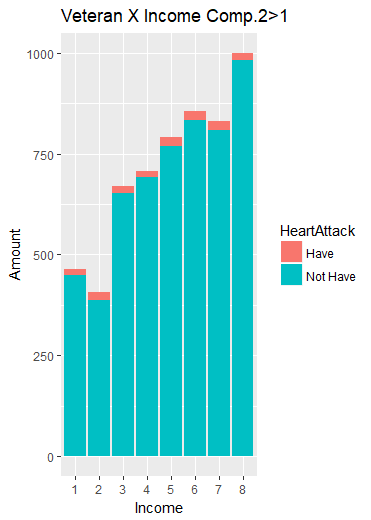
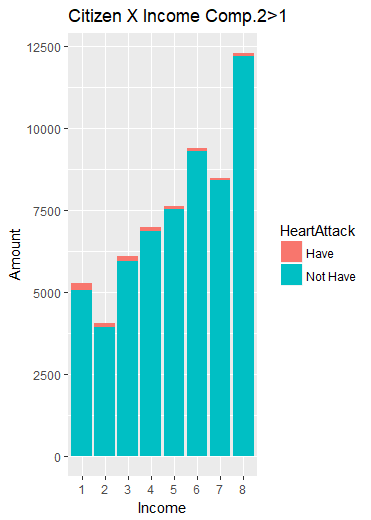
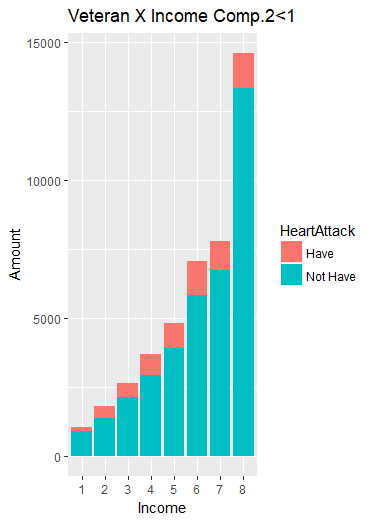
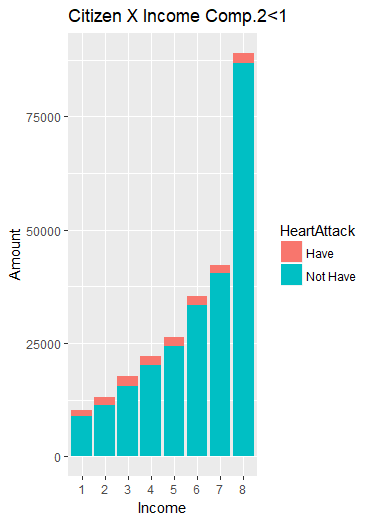
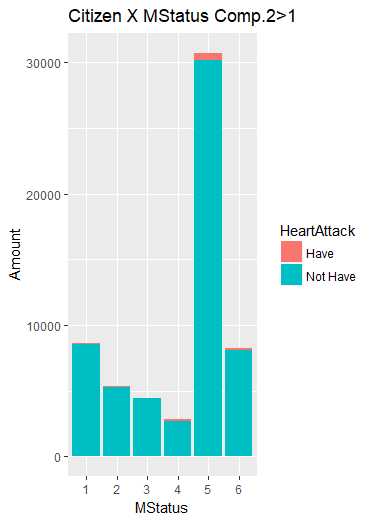
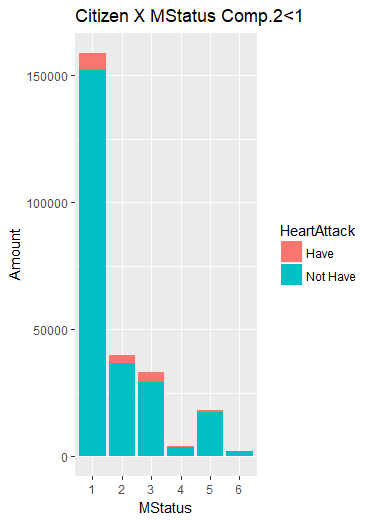


Figure 18: The distribution on Income for the group with C2<1 (left) and for the group with C2>1(right).



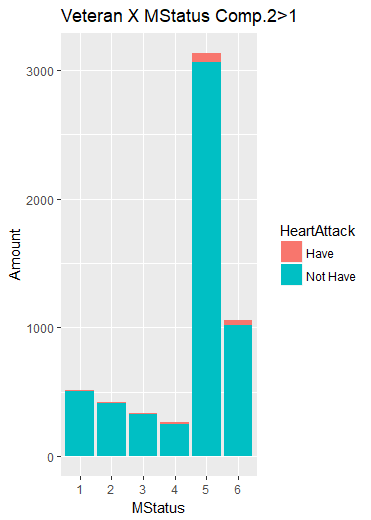
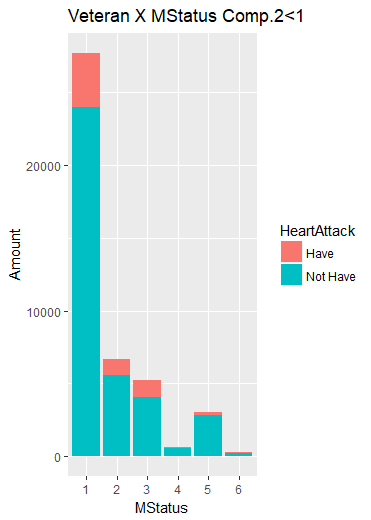
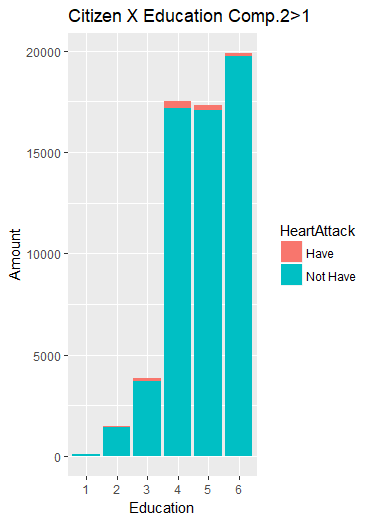
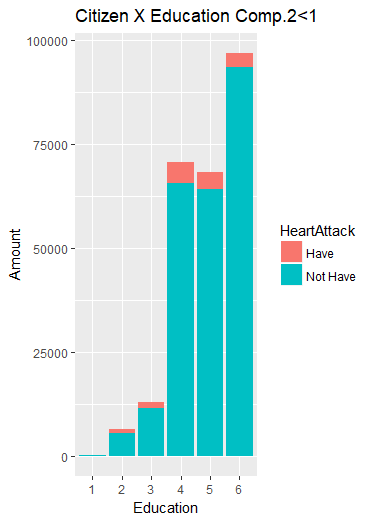


Figure 19: The distribution on Marital Status for the group with C2<1 (left) and for the group with C2>1(right).

Figure 18 and Figure 19 show that Income and Marital Status distributed differently in area of C2<1 and C2>1 for either of Citizens and Veterans. In the area of C2<1, MStatus level 1 is the highest incidence rate level for both of Veterans (13.5%) and Citizens (4.3%). Additionally, the distributions for Citizens and Veterans are very similar in each area (all p-values larger than 0.39).



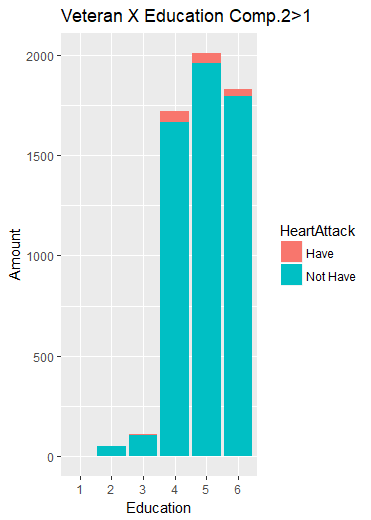
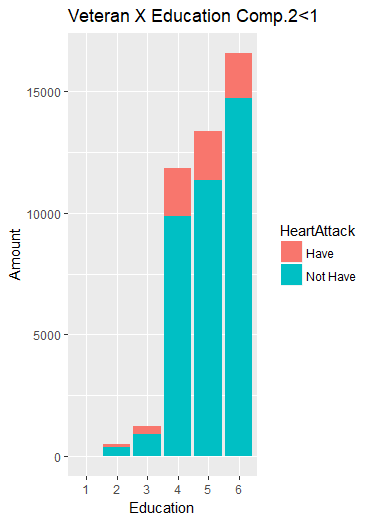


Figure 20: The distribution on Education for the group with C2<1 (left) and for the group with C2>1(right).

The difference between the Education distributions in two areas is not as large as other factors. It may be because that Education level 4 to 6 take a large part of the whole Data. However, the incidence rates of level 4 to 6 for Veteran are much higher than those for Citizens.

According to the details of Component 2, C2<1 represents one or more characteristics among high Age, high Income, female, married, and unemployed or retired.

On the other hand, comparing the two plots, blue units are distributed differently. In veteran plot, there are more blue units in upper area (positive C3 value), while the blue units in citizen plot are almost evenly distributed from upper to lower. A line of C3=0 is added on the plots due to visually analysis.

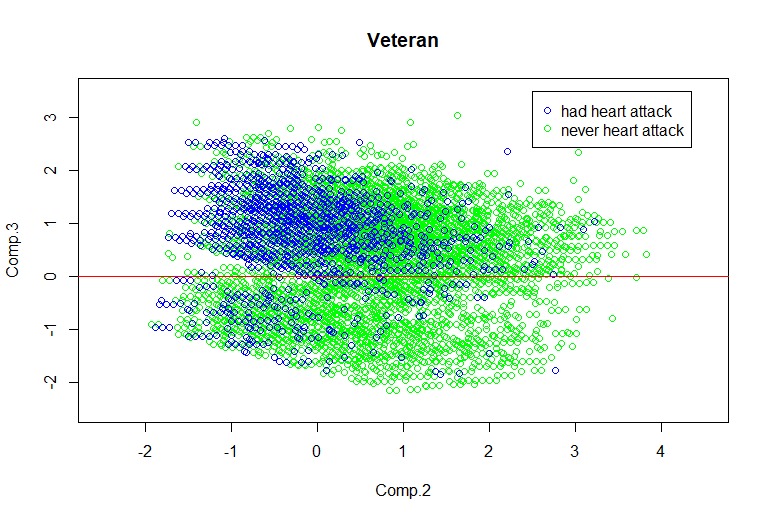
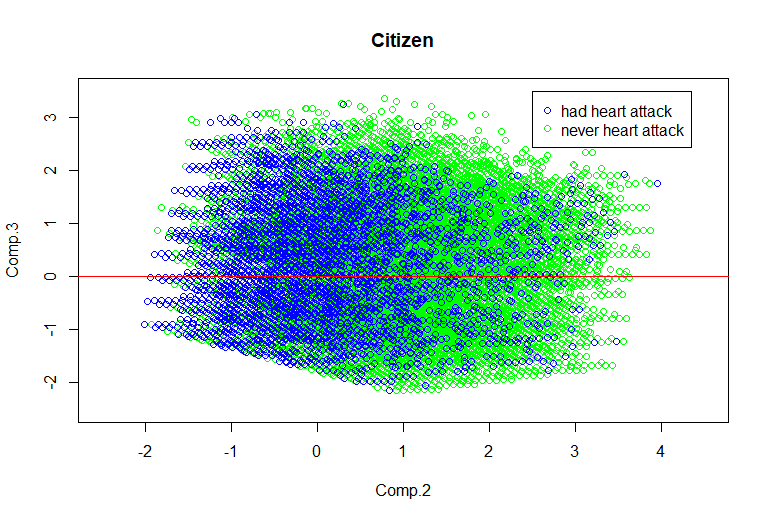
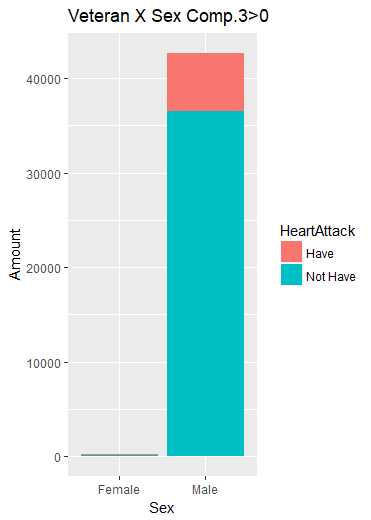
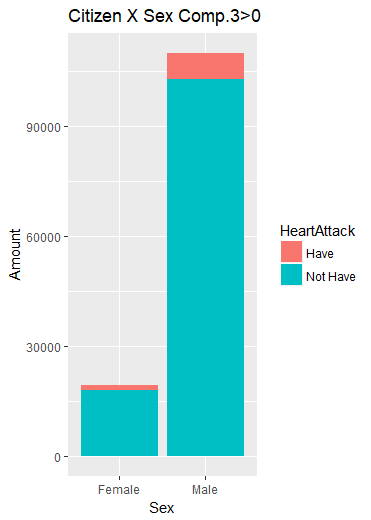
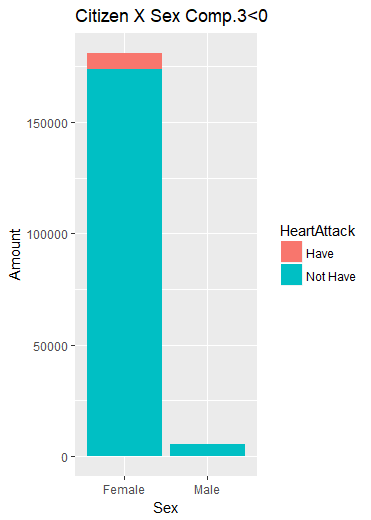


Figure 21: The distribution on Education for the group with C2<1 (left) and for the group with C2>1(right)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Comp.3 | C3<0 | | C3>0 | |
|  | Veteran | Citizen | Veteran | Citizen |
| Heart Attack=1(Yes) | 250(3.9%) | 7327(3.9%) | 6218(14.5%) | 8708(6.7%) |
| Heart Attack=2(No) | 6097(96.1%) | 179207(96.1%) | 36657(85.5%) | 120851(93.3%) |

Table 12: the incidence rate for Veterans and Citizens in the area of C3<0 and C3>0.

In the area of C3<0, 3.9% citizen units and 3.9% veteran units are patients, while there is still a large difference in the incidence rate between two groups in the area of C3>0. Finding out the similarity of two groups in the area C3<0 is helpful for analyzing the reason of the higher incidence rate for Veterans.



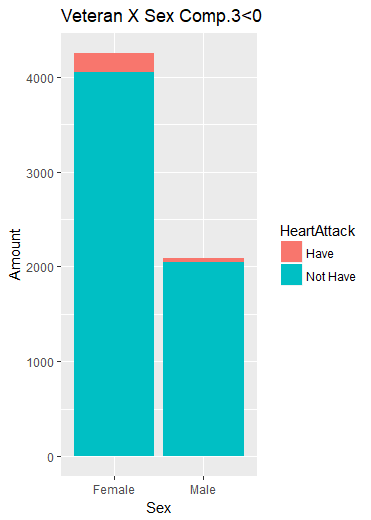
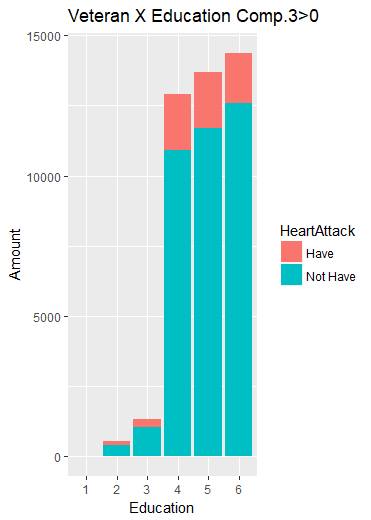
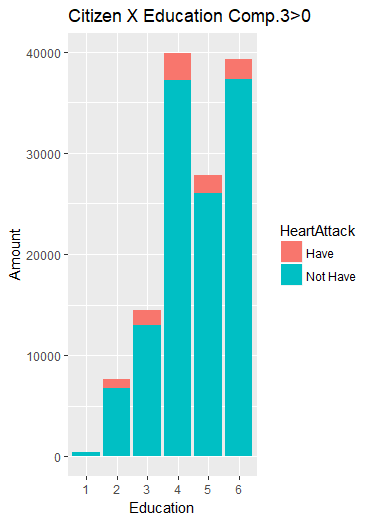


Figure 22: The distribution on Sex for the group with C3<0 (left) and for the group with C3>0(right)

Different with the whole data, gender is distributed more similar for two groups in C3<0 area and C3>0 area. Female units are dominate in C3<0, while Male units are dominate in C3>0. The difference between the two areas could be explained as a difference between two Genders.



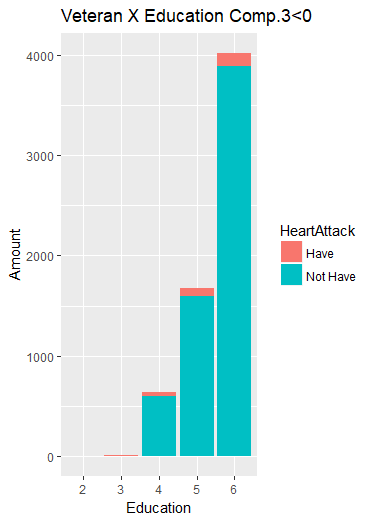
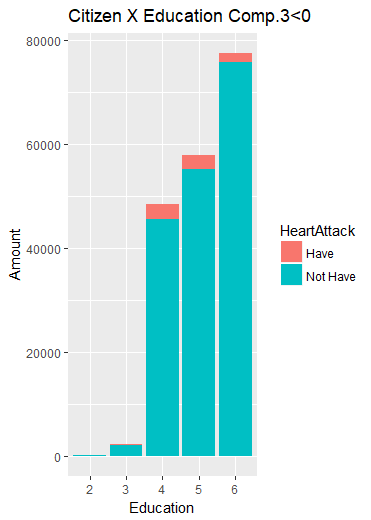
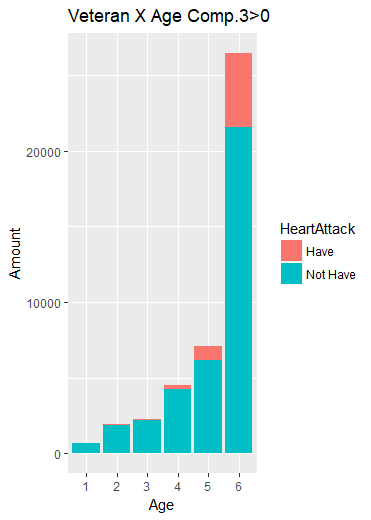
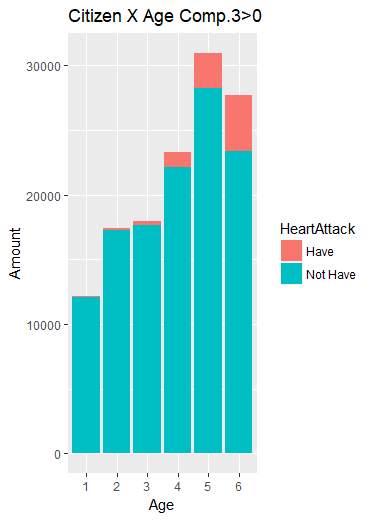


Figure 23: The distribution on Education for the group with C3<0 (left) and for the group with C3>0(right)

Education levels distributions are still complicated for each group and each area. However, the incidence rate of Veterans is higher than Citizens in C3>0 area.



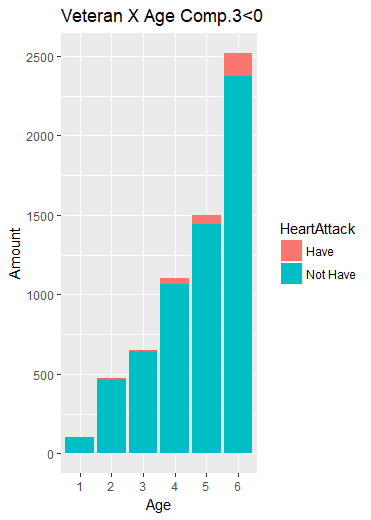
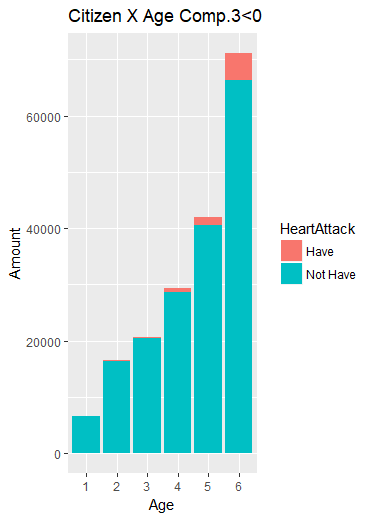


Figure 24: The distribution on Age for the group with C3<0 (left) and for the group with C3>0(right)

For the area of C3<0, the Age distributions of Citizens and Veterans are almost the same (p-value=0.99), and the incident rate for each Age level are almost same (p-value=0.62), too. For the area of C3>0, the incident rates of Age level 6 are 15.6% for Citizens and 18.7% for Veterans. Both of them are very high.

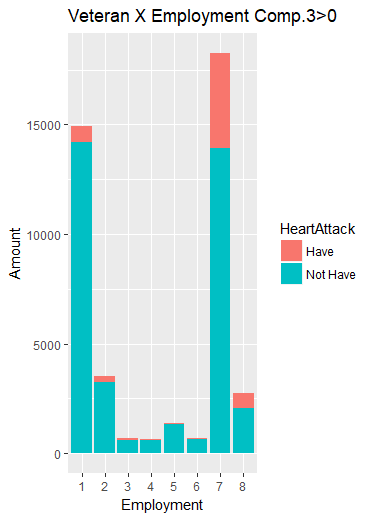
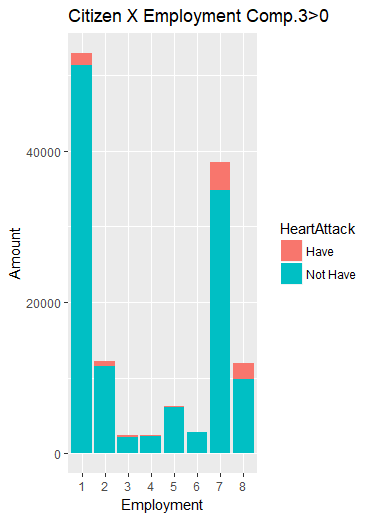
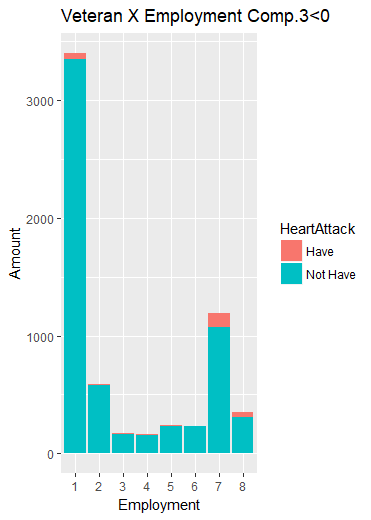
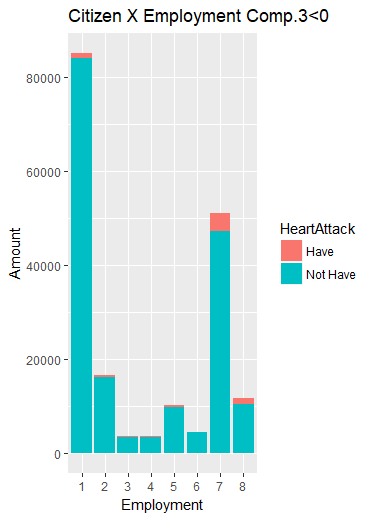
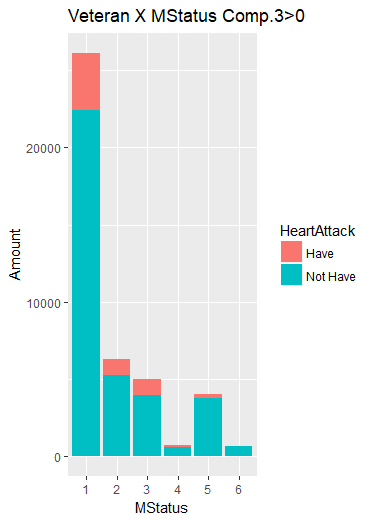
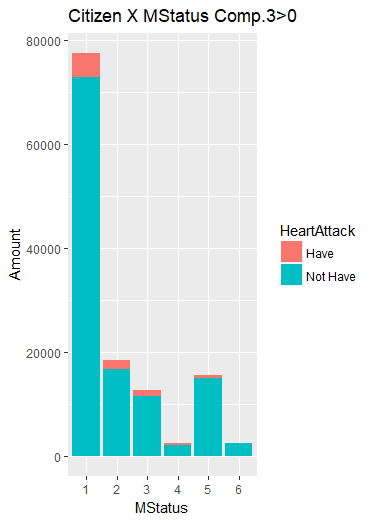


Figure 25: The distribution on Employment for the group with C3<0 (left) and for the group with C3>0(right)

The two groups have slightly different distributions in the area of C3>0 due to the Employment level 7. For Veteran group units in C3>0, Employment level 7 takes 42.6% of them with an incidence rate of 0.239, while it is only 29.8% for Citizen units with an incidence rate of 0.097.



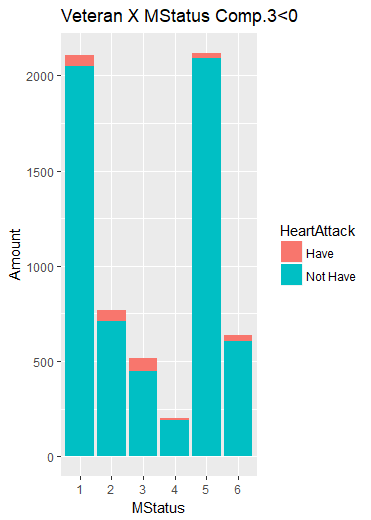
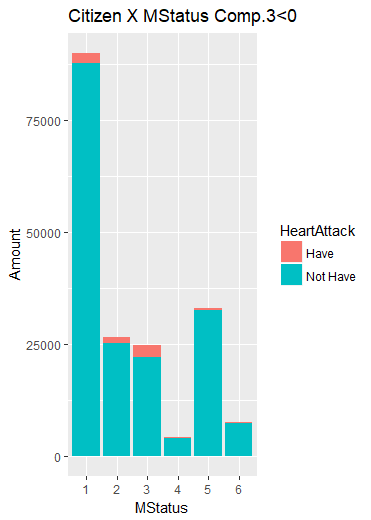
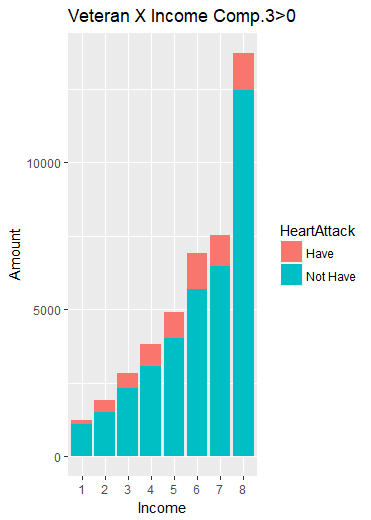
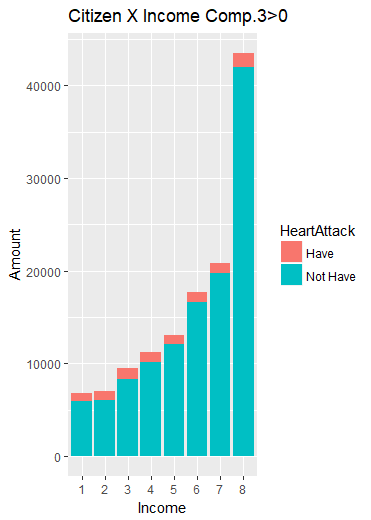


Figure 26: The distribution on Marital Status for the group with C3<0 (left) and for the group with C3>0(right)



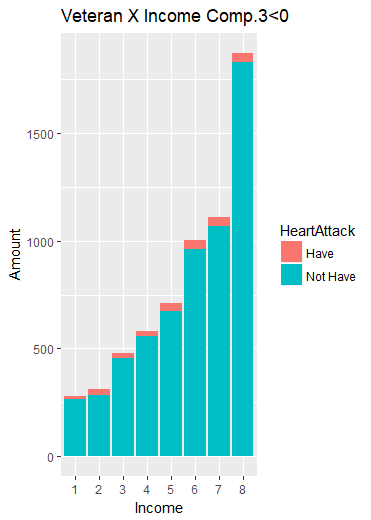
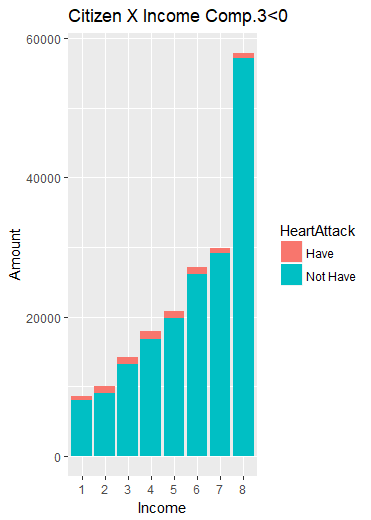


Figure 27: The distribution on Income for the group with C3<0 (left) and for the group with C3>0(right)

Even though the two groups have similar distribution on MStatus (p-value=0.77) and Income (p-value=0.55) in the area of C3>0, the incidence rates of Veterans are higher than Citizens, especially in MStatus level 1.

Based on the PCA results, Veteran experience and Sex do have a strong interaction with the factors including Employment, Marital Status, Income and Age. These interactions should be included in the logistic model.

**4-3 the Logistic Model**

Since Principal Component Analysis provides the potential relationship between several interaction and the incidence rate of Heart Attack, applying this model in logistic regression with original factors could do help in detecting the probability of Heart Attack for each single unit.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Df | Deviance | Resid. Df | Resid. Dev | Pr(>Chi) |
| NULL | 1 |  | 255719 | 118357 | < 2.2e-16 \*\*\* |
| Veteran | 1 | 2817.5 | 255718 | 115539 | < 2.2e-16 \*\*\* |
| Sex | 1 | 563.5 | 255717 | 114975 | < 2.2e-16 \*\*\* |
| Education | 5 | 1722.6 | 255712 | 113253 | < 2.2e-16 \*\*\* |
| Income | 7 | 3141.7 | 255705 | 110111 | < 2.2e-16 \*\*\* |
| Age | 5 | 7538.8 | 255700 | 102572 | < 2.2e-16 \*\*\* |
| Employment | 7 | 4280.6 | 255693 | 98292 | < 2.2e-16 \*\*\* |
| MStatus | 5 | 451.1 | 255688 | 97841 | < 2.2e-16 \*\*\* |
| Veteran:Income | 7 | 151.3 | 255681 | 97689 | < 2.2e-16 \*\*\* |
| Sex:Income | 7 | 237.8 | 255674 | 97452 | < 2.2e-16 \*\*\* |
| Veteran:Age | 5 | 17.0 | 255669 | 97435 | 0.004477 \*\* |
| Sex:Age | 5 | 61.7 | 255664 | 97373 | 5.425e-12 \*\*\* |
| Veteran:Employment | 7 | 273.5 | 255657 | 97099 | < 2.2e-16 \*\*\* |
| Sex:Employment | 7 | 452.5 | 255650 | 96647 | < 2.2e-16 \*\*\* |
| Veteran:MStatus | 5 | 27.8 | 255645 | 96619 | 3.979e-05 \*\*\* |
| Sex:MStatus | 5 | 208.7 | 255640 | 96410 | < 2.2e-16 \*\*\* |

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Table 13: The anova table for logistic regression model.

**5. Discussion**

In the raw data, a huge difference between veterans and citizens in heart attack incidence rates is shown. But the difference is much smaller in Rubin-Casual model. It is worthy to discuss in the future that whether this 1% difference indicates the incidence rates of veterans and citizens are different in real world.

From the logistic regression model, the predicted values are among 0.4 to 0.999, when 0 represents of had Heart Attack before and 1 represents of not. Around 93% of units’ Heart Attack values are 1 in the data, thus the predicted values are more close to 1. It is important to set the boundary value to judge the incidence based on the predicted value. When the value of boundary is changing from low to high, more Heart Attack cases are detected successfully, but more healthy people are detected as patients wrongly.

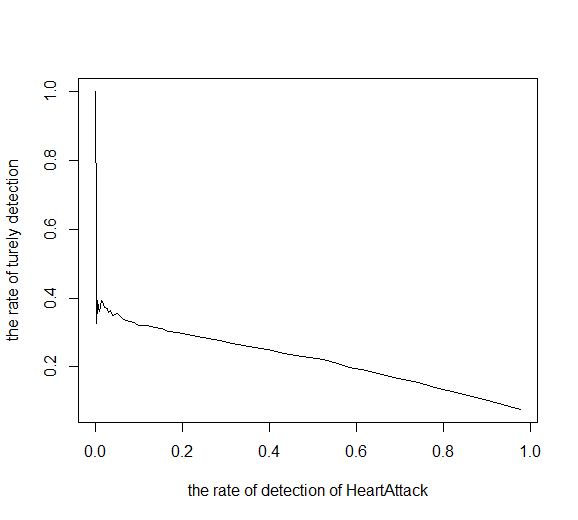


Figure 28：The accuracy of the detection.

**6. References**

Blocks：2(Sex)\*6(Education)\*6(Age)\*6(Marital Status)\*8(Employment)\*8(Income) =27648. However, only 15873 blocks are not empty.

Data: There are 316,093 citizens and 49,222 veterans.

Method 1:

Within each block, the same and fix amount of citizens and veterans were randomly selected. The selecting amount in each block was set to be as large as possible. Finally, 42,161 citizen and 42,161 veterans were selected. The propensity scores of the two groups are the same due to the matching is the perfect matching (same distributions on every variable). After repeating the random selection with random seeds of 1 to 80, the Heart Attack incidences of the two groups were obtained. Among the 80 times repeating, veterans’ incidences were always higher than citizens’ around 0.01 (mean incidence 0.106 for citizen and 0.118 for veterans), which is significant in proportion test.

In this method, the selected data structure is more like the structure of veterans, since more than 85% of veterans were selected.

Method2:

For generalize the result,

The incidences for Veterans and Citizens in each block were calculated first if there were 4 types of units (veteran or not, had Heart Attack or not) in this block. If the block does not include any of them, the block would be ignored. Only 796 blocks left. Then, the weight of each block for Citizens was calculated based on the number of citizens in each block and the total amount of citizens (excludes those ignored blocks). Last, the weighted average incidences of veterans and citizens were calculated, both of them based on the block weight of citizens.

After weighing the incidence in each block, the veterans’ Heart Attack incidence is 0.158806 and the citizens’ is 0.094979. Apparently, the weighted incidences of Citizen and Veteran are slightly higher than the original incidences (0.0507 for Citizen and 0.1314 for Veteran), while the difference between them is smaller slightly. But the difference between the incidences is still extremely significant.

**0.06932686**

**0.130272**